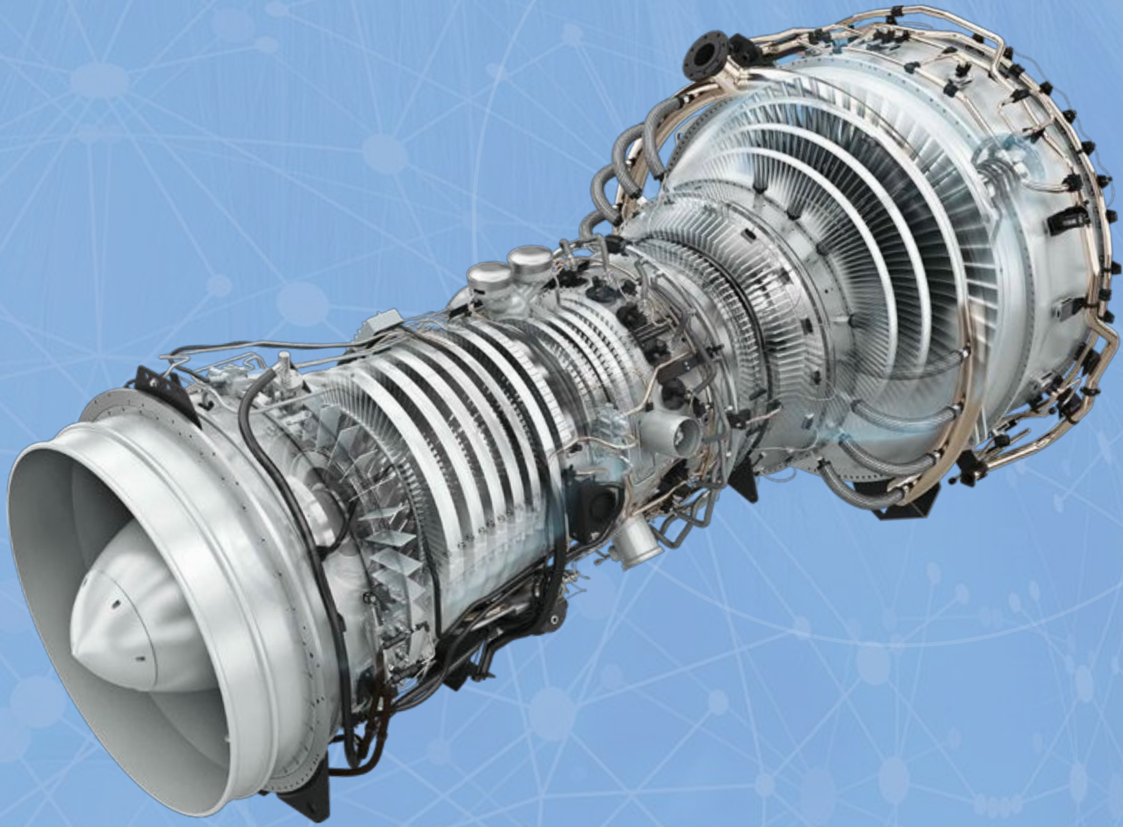
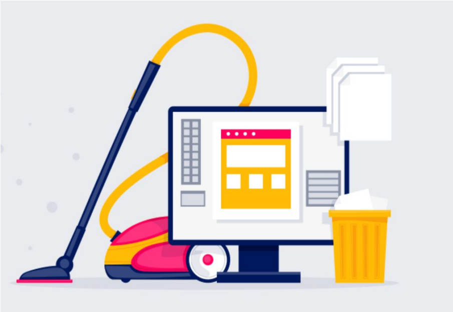
Gas Turbine Predictive

Team: Predictive Lions



Methodology

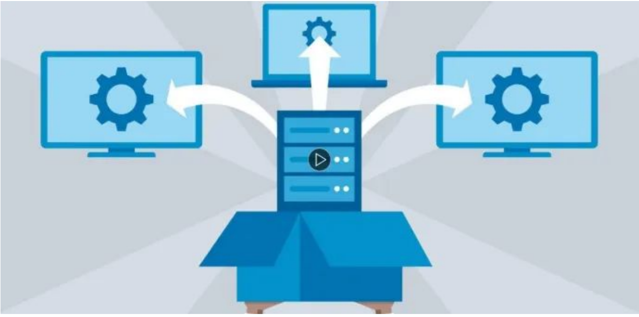
Data Exploration



Data Cleansing and Feature Engineering

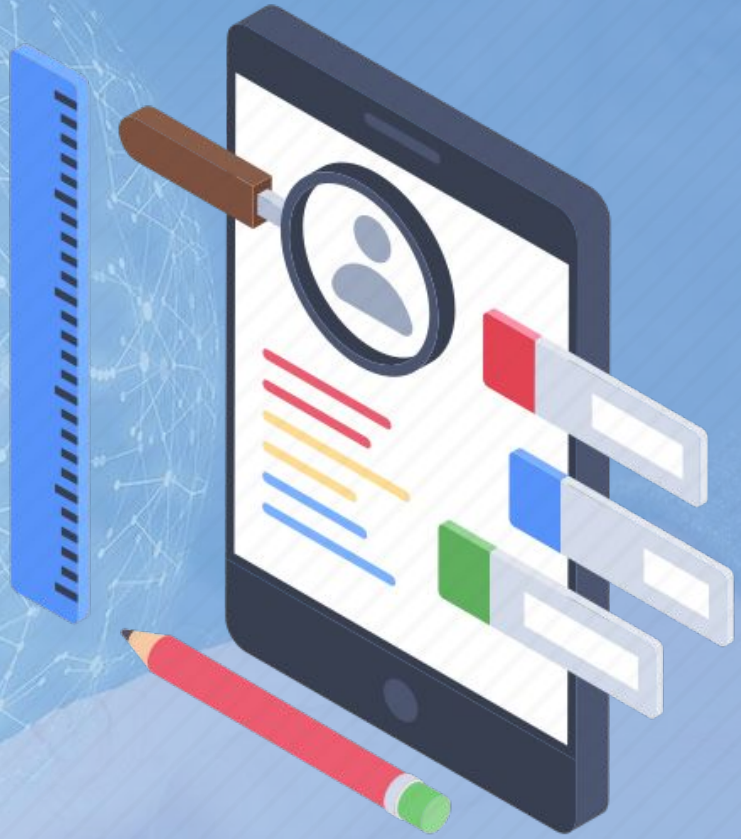
Model Selection and Comparison

Selected Model Evaluation

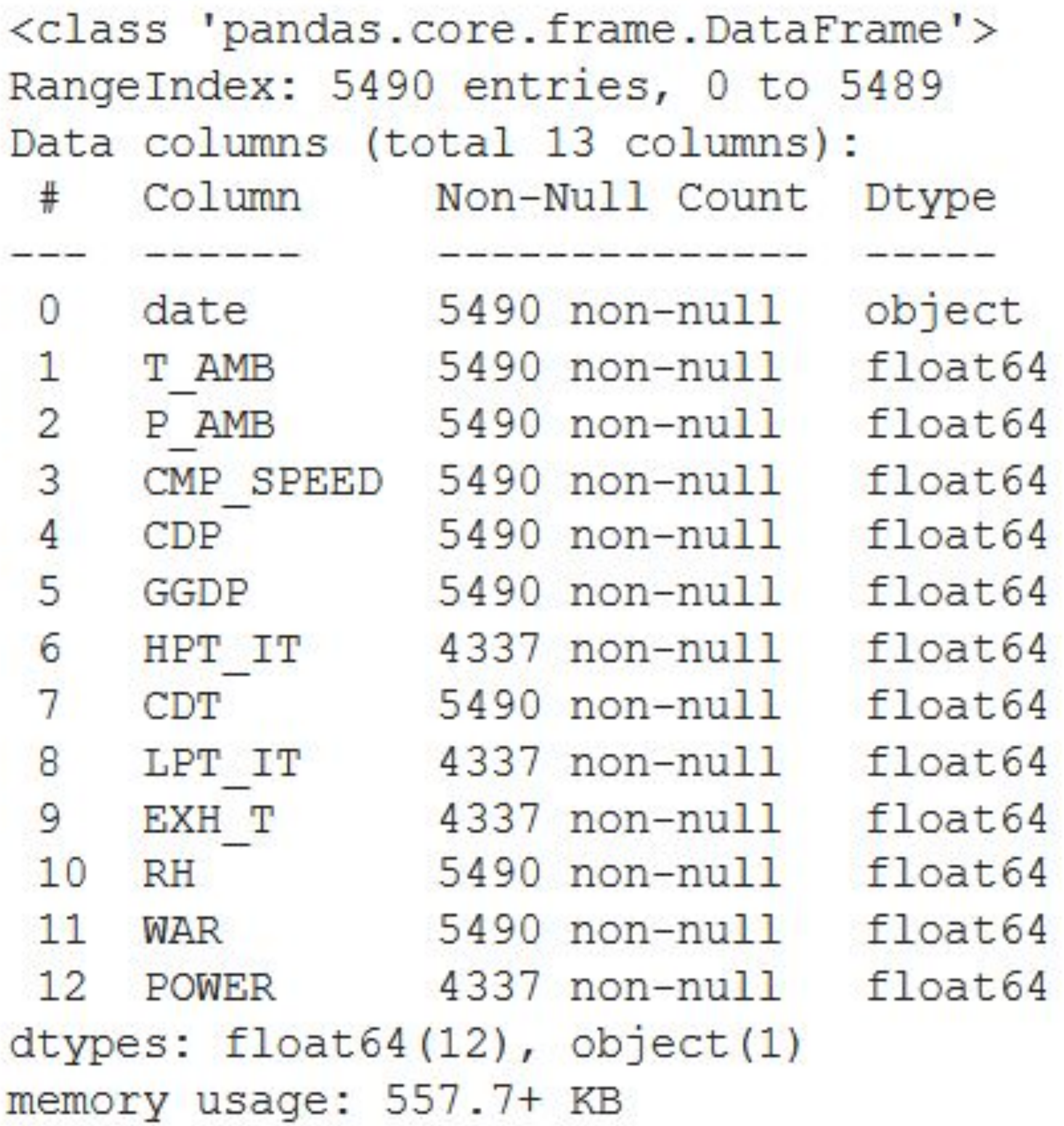


Conclusions

DATA EXPLORATION



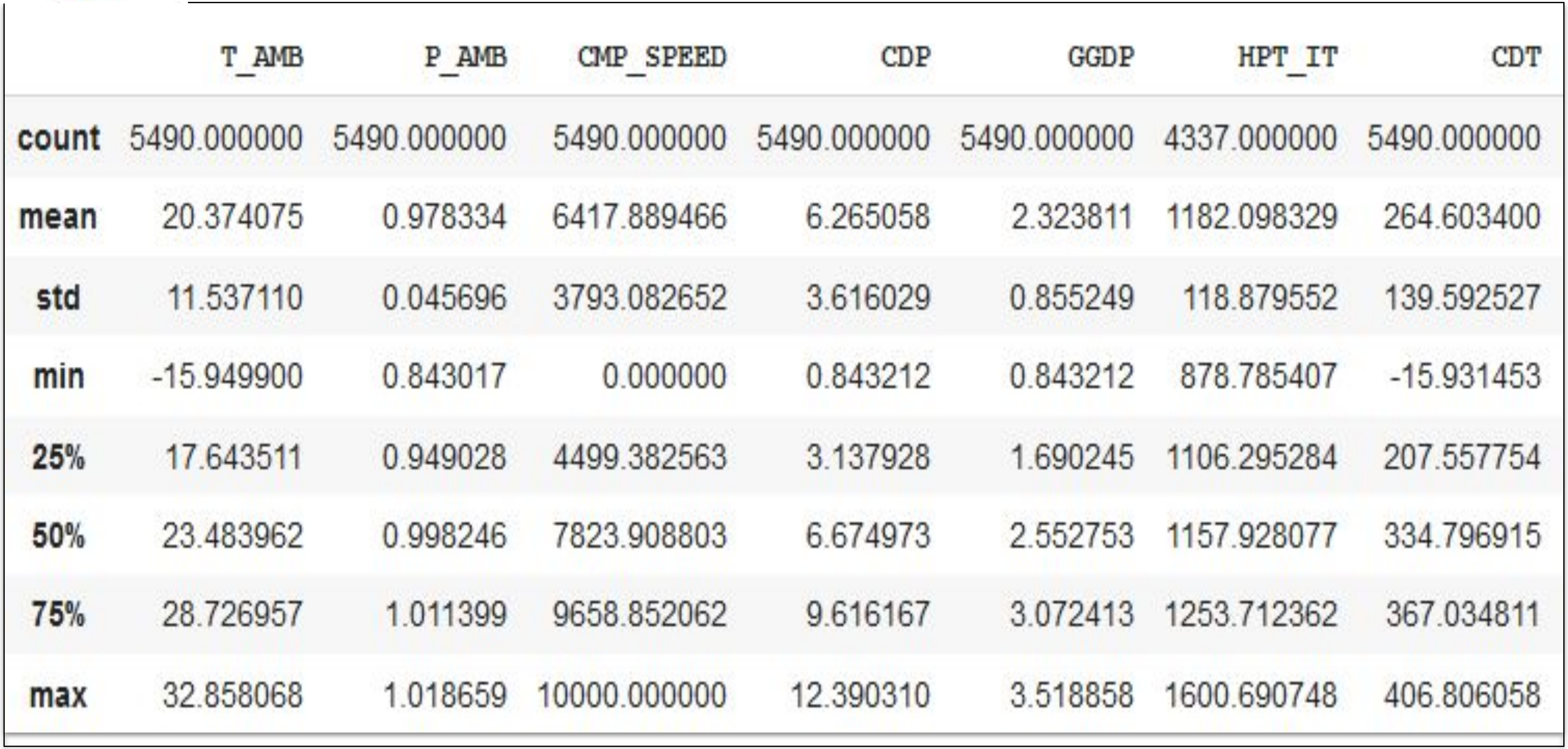
Information of the given dataset



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5490 entries, 0 to 5489
Data columns (total 13 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   date       5490 non-null   object
 1   T_AMB      5490 non-null   float64
 2   P_AMB      5490 non-null   float64
 3   CMP_SPEED  5490 non-null   float64
 4   CDP        5490 non-null   float64
 5   GGDP       5490 non-null   float64
 6   HPT_IT     4337 non-null   float64
 7   CDT        5490 non-null   float64
 8   LPT_IT     4337 non-null   float64
 9   EXH_T      4337 non-null   float64
 10  RH         5490 non-null   float64
 11  WAR        5490 non-null   float64
 12  POWER      4337 non-null   float64
dtypes: float64(12), object(1)
memory usage: 557.7+ KB
```

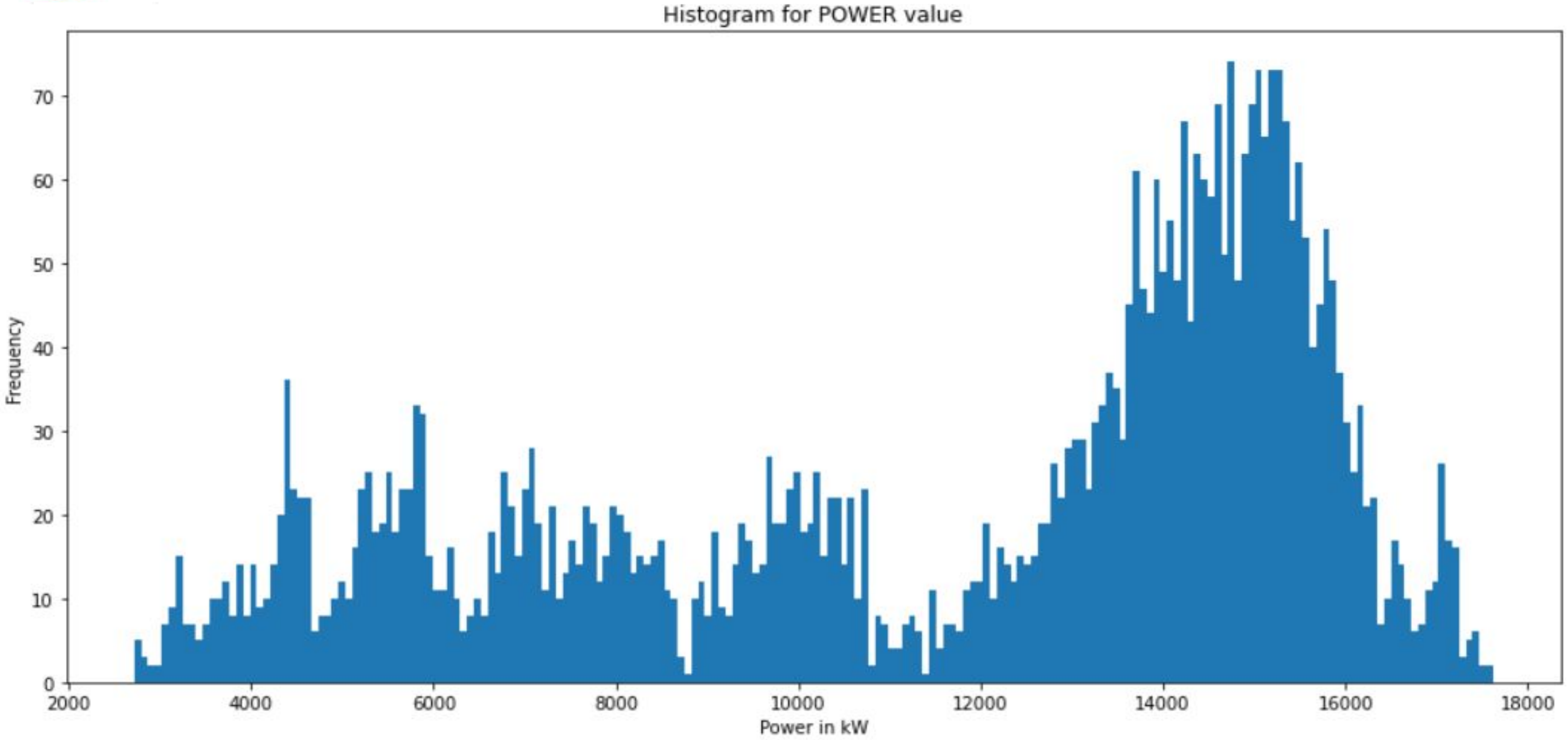
As a first step, we verify the characteristics of the volume of data obtained.

Basic Statistical Indicators



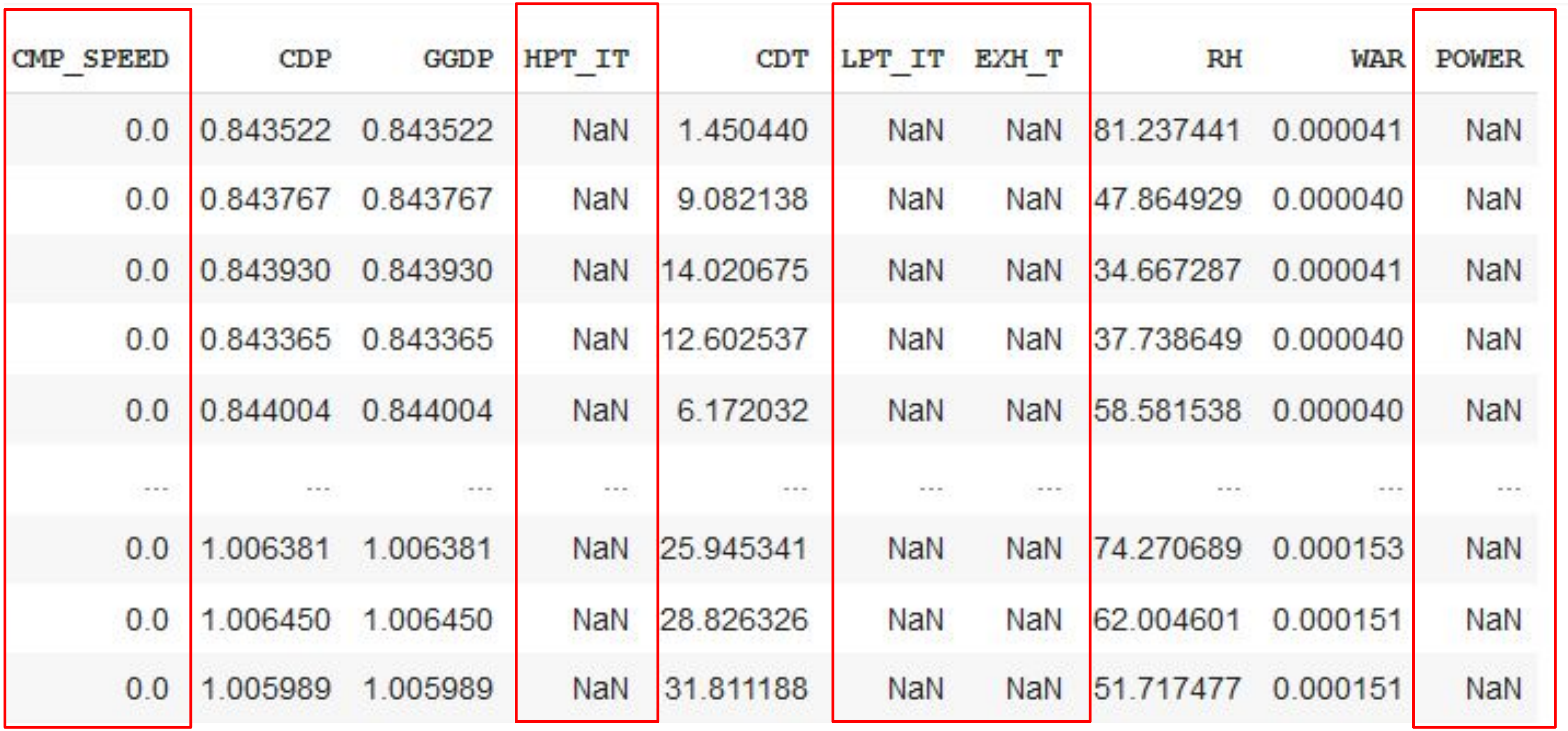
	T_AMB	P_AMB	CMP_SPEED	CDP	GGDP	HPT_IT	CDT
count	5490.000000	5490.000000	5490.000000	5490.000000	5490.000000	4337.000000	5490.000000
mean	20.374075	0.978334	6417.889466	6.265058	2.323811	1182.098329	264.603400
std	11.537110	0.045696	3793.082652	3.616029	0.855249	118.879552	139.592527
min	-15.949900	0.843017	0.000000	0.843212	0.843212	878.785407	-15.931453
25%	17.643511	0.949028	4499.382563	3.137928	1.690245	1106.295284	207.557754
50%	23.483962	0.998246	7823.908803	6.674973	2.552753	1157.928077	334.796915
75%	28.726957	1.011399	9658.852062	9.616167	3.072413	1253.712362	367.034811
max	32.858068	1.018659	10000.000000	12.390310	3.518858	1600.690748	406.806058

Frequency distribution of the dependent variable POWER



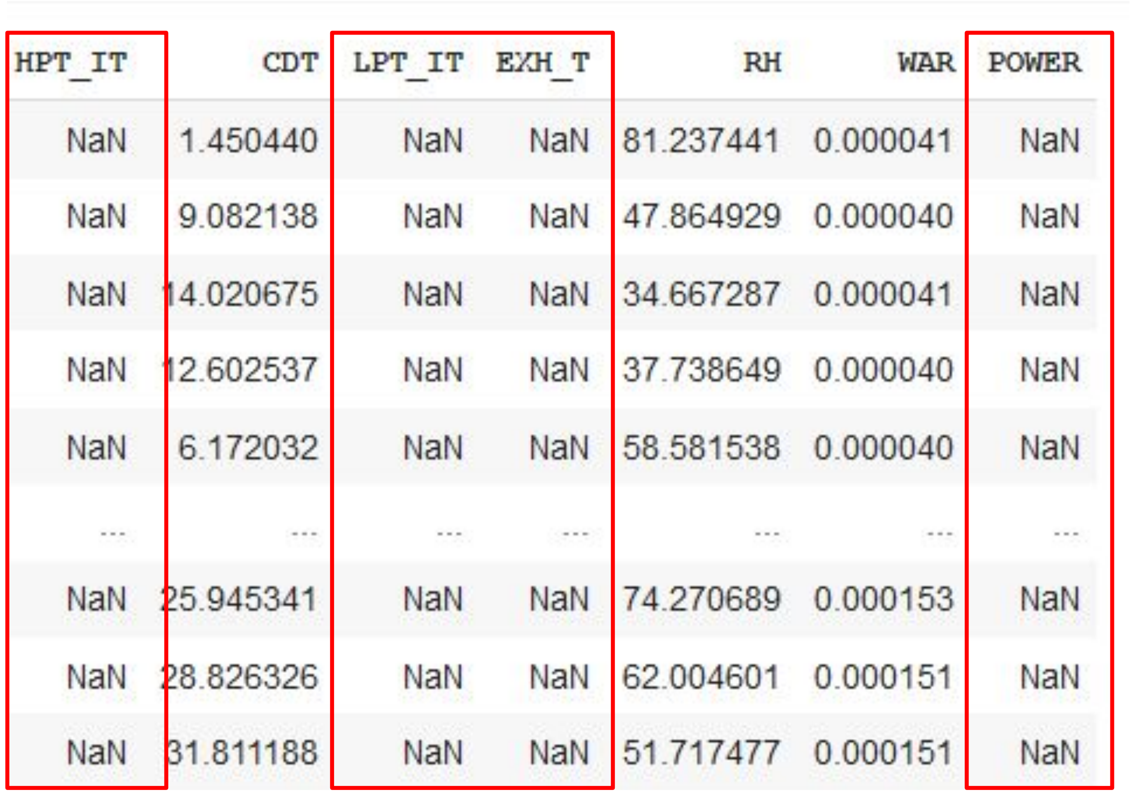
DATA CLEANSING AND FEATURE ENGINEERING

Null value analysis

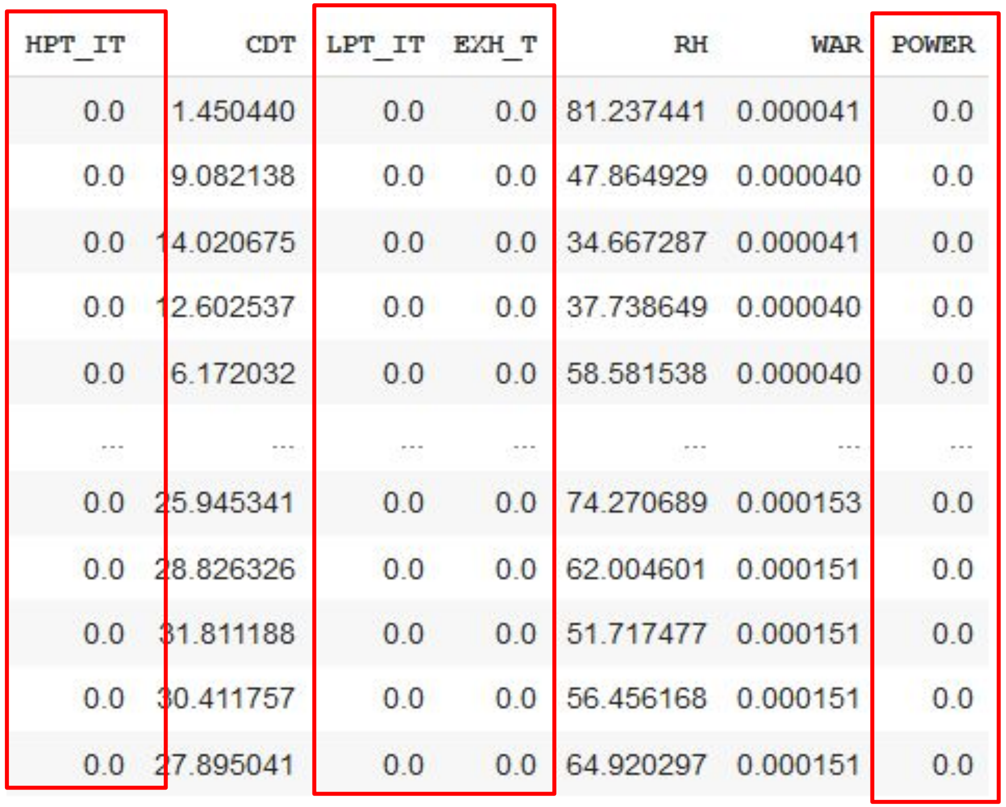


CMP_SPEED	CDP	GGDP	HPT_IT	CDT	LPT_IT	EXH_T	RH	WAR	POWER
0.0	0.843522	0.843522	NaN	1.450440	NaN	NaN	81.237441	0.000041	NaN
0.0	0.843767	0.843767	NaN	9.082138	NaN	NaN	47.864929	0.000040	NaN
0.0	0.843930	0.843930	NaN	14.020675	NaN	NaN	34.667287	0.000041	NaN
0.0	0.843365	0.843365	NaN	12.602537	NaN	NaN	37.738649	0.000040	NaN
0.0	0.844004	0.844004	NaN	6.172032	NaN	NaN	58.581538	0.000040	NaN
...	...	...	...	...	...	...	...	...	...
0.0	1.006381	1.006381	NaN	25.945341	NaN	NaN	74.270689	0.000153	NaN
0.0	1.006450	1.006450	NaN	28.826326	NaN	NaN	62.004601	0.000151	NaN
0.0	1.005989	1.005989	NaN	31.811188	NaN	NaN	51.717477	0.000151	NaN

Imputation



HPT_IT	CDT	LPT_IT	EXH_T	RH	WAR	POWER
NaN	1.450440	NaN	NaN	81.237441	0.000041	NaN
NaN	9.082138	NaN	NaN	47.864929	0.000040	NaN
NaN	14.020675	NaN	NaN	34.667287	0.000041	NaN
NaN	12.602537	NaN	NaN	37.738649	0.000040	NaN
NaN	6.172032	NaN	NaN	58.581538	0.000040	NaN
...	...	...	...	...	...	...
NaN	25.945341	NaN	NaN	74.270689	0.000153	NaN
NaN	28.826326	NaN	NaN	62.004601	0.000151	NaN
NaN	31.811188	NaN	NaN	51.717477	0.000151	NaN

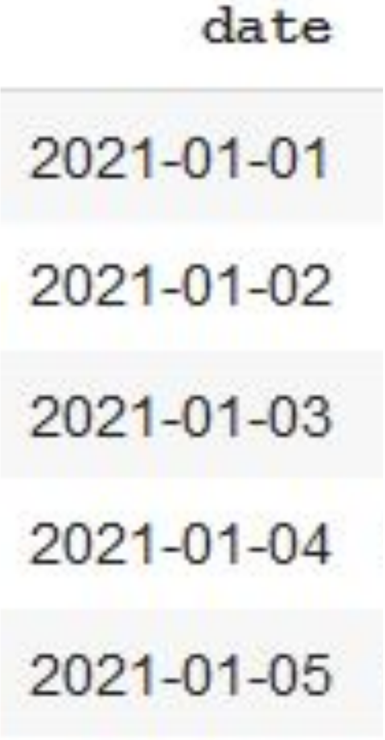


HPT_IT	CDT	LPT_IT	EXH_T	RH	WAR	POWER
0.0	1.450440	0.0	0.0	81.237441	0.000041	0.0
0.0	9.082138	0.0	0.0	47.864929	0.000040	0.0
0.0	14.020675	0.0	0.0	34.667287	0.000041	0.0
0.0	12.602537	0.0	0.0	37.738649	0.000040	0.0
0.0	6.172032	0.0	0.0	58.581538	0.000040	0.0
...	...	...	...	...	...	...
0.0	25.945341	0.0	0.0	74.270689	0.000153	0.0
0.0	28.826326	0.0	0.0	62.004601	0.000151	0.0
0.0	31.811188	0.0	0.0	51.717477	0.000151	0.0
0.0	30.411757	0.0	0.0	56.456168	0.000151	0.0
0.0	27.895041	0.0	0.0	64.920297	0.000151	0.0

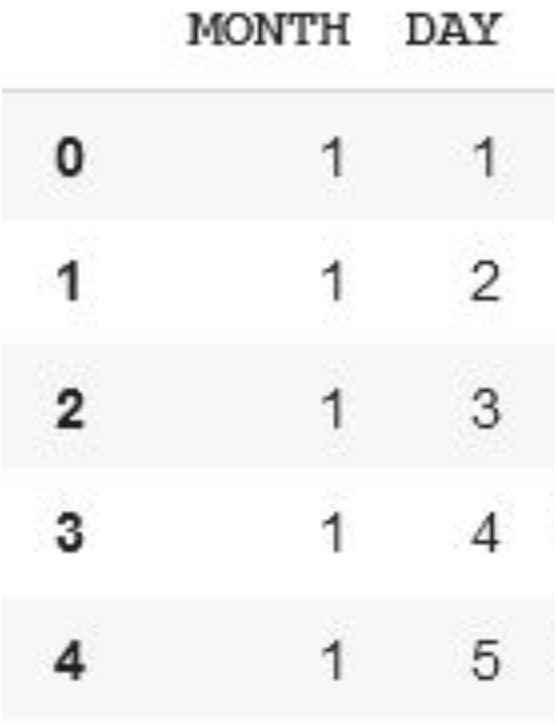
Data extraction

How can we process the date without losing information?

Could the season of the year affect our target?

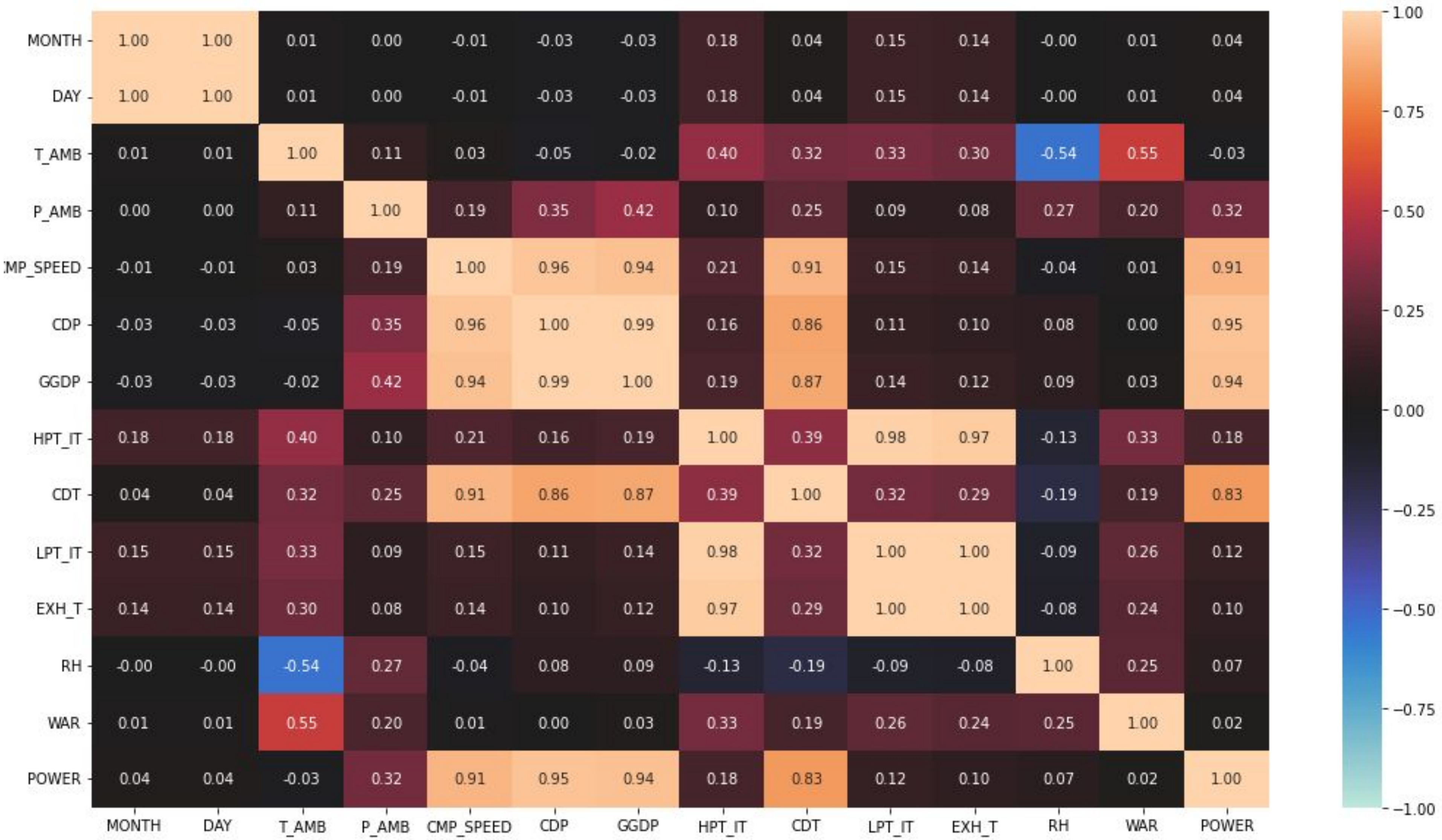


date
2021-01-01
2021-01-02
2021-01-03
2021-01-04
2021-01-05



	MONTH	DAY
0	1	1
1	1	2
2	1	3
3	1	4
4	1	5

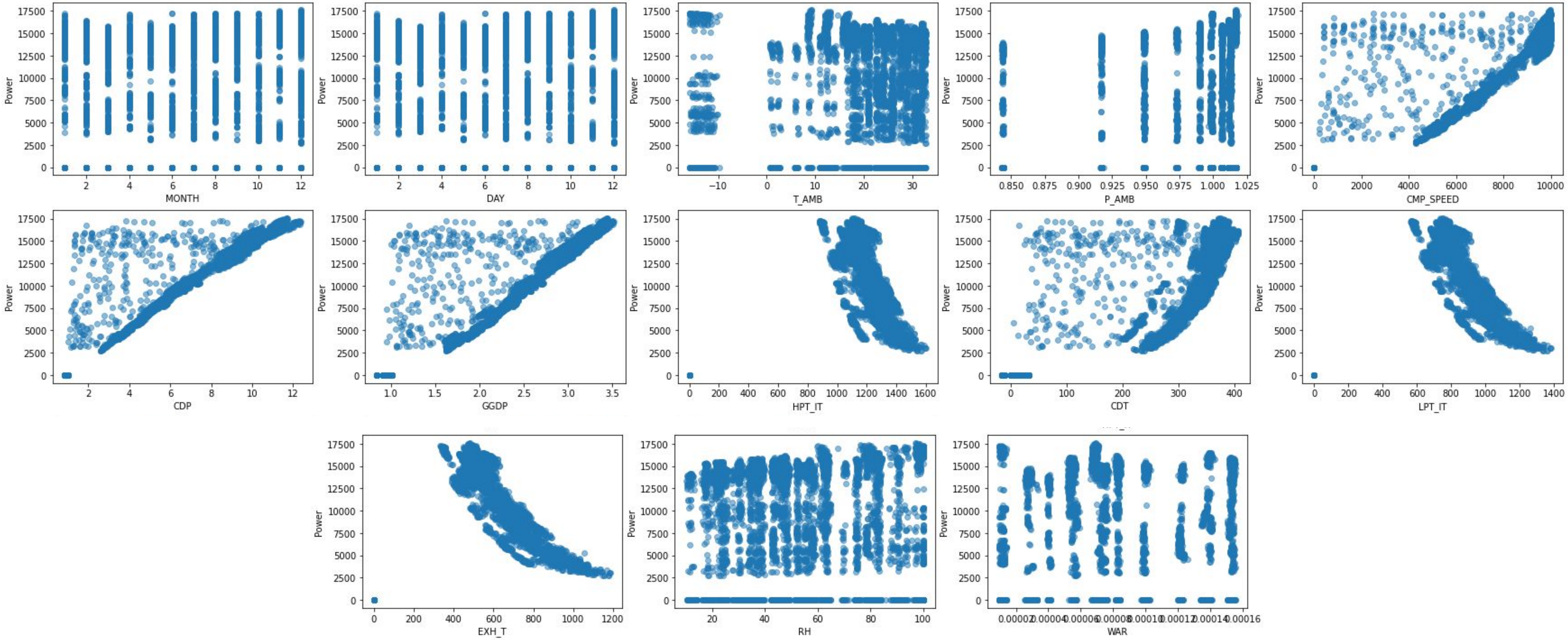
Feature selection -Correlation of variables



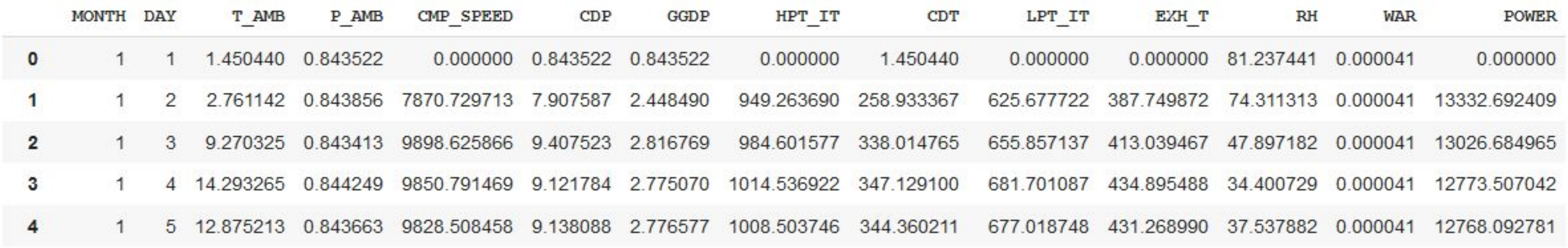
	MONTH	DAY	T_AMB	P_AMB	CMP_SPEED	CDP	GGDP	HPT_IT	CDT	LPT_IT	EXH_T	RH	WAR	POWER
MONTH	1.00	1.00	0.01	0.00	-0.01	-0.03	-0.03	0.18	0.04	0.15	0.14	-0.00	0.01	0.04
DAY	1.00	1.00	0.01	0.00	-0.01	-0.03	-0.03	0.18	0.04	0.15	0.14	-0.00	0.01	0.04
T_AMB	0.01	0.01	1.00	0.11	0.03	-0.05	-0.02	0.40	0.32	0.33	0.30	-0.54	0.55	-0.03
P_AMB	0.00	0.00	0.11	1.00	0.19	0.35	0.42	0.10	0.25	0.09	0.08	0.27	0.20	0.32
MP_SPEED	-0.01	-0.01	0.03	0.19	1.00	0.96	0.94	0.21	0.91	0.15	0.14	-0.04	0.01	0.91
CDP	-0.03	-0.03	-0.05	0.35	0.96	1.00	0.99	0.16	0.86	0.11	0.10	0.08	0.00	0.95
GGDP	-0.03	-0.03	-0.02	0.42	0.94	0.99	1.00	0.19	0.87	0.14	0.12	0.09	0.03	0.94
HPT_IT	0.18	0.18	0.40	0.10	0.21	0.16	0.19	1.00	0.39	0.98	0.97	-0.13	0.33	0.18
CDT	0.04	0.04	0.32	0.25	0.91	0.86	0.87	0.39	1.00	0.32	0.29	-0.19	0.19	0.83
LPT_IT	0.15	0.15	0.33	0.09	0.15	0.11	0.14	0.98	0.32	1.00	1.00	-0.09	0.26	0.12
EXH_T	0.14	0.14	0.30	0.08	0.14	0.10	0.12	0.97	0.29	1.00	1.00	-0.08	0.24	0.10
RH	-0.00	-0.00	-0.54	0.27	-0.04	0.08	0.09	-0.13	-0.19	-0.09	-0.08	1.00	0.25	0.07
WAR	0.01	0.01	0.55	0.20	0.01	0.00	0.03	0.33	0.19	0.26	0.24	0.25	1.00	0.02
POWER	0.04	0.04	-0.03	0.32	0.91	0.95	0.94	0.18	0.83	0.12	0.10	0.07	0.02	1.00

Correlation Feature > |0.1|

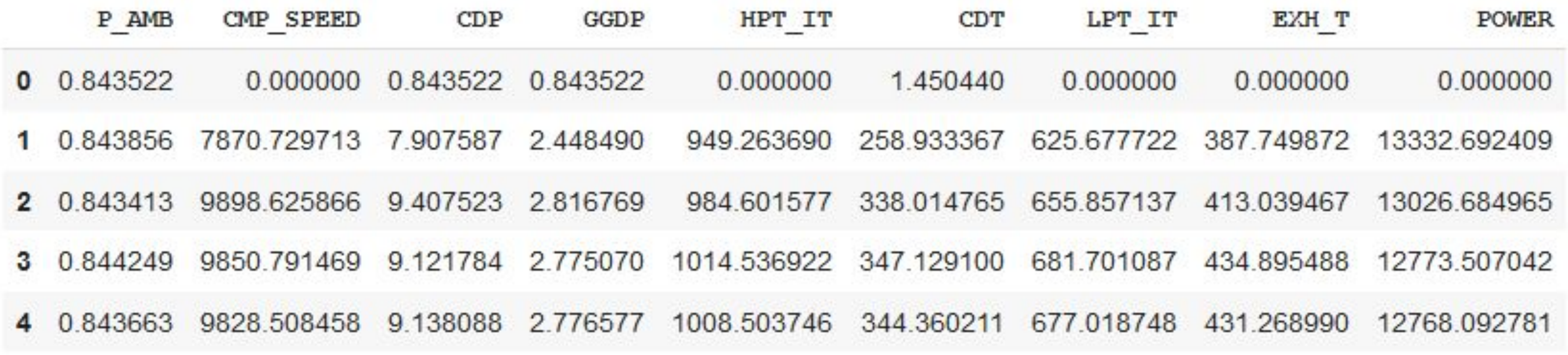
Feature selection - Correlation justification



Feature split / dropping



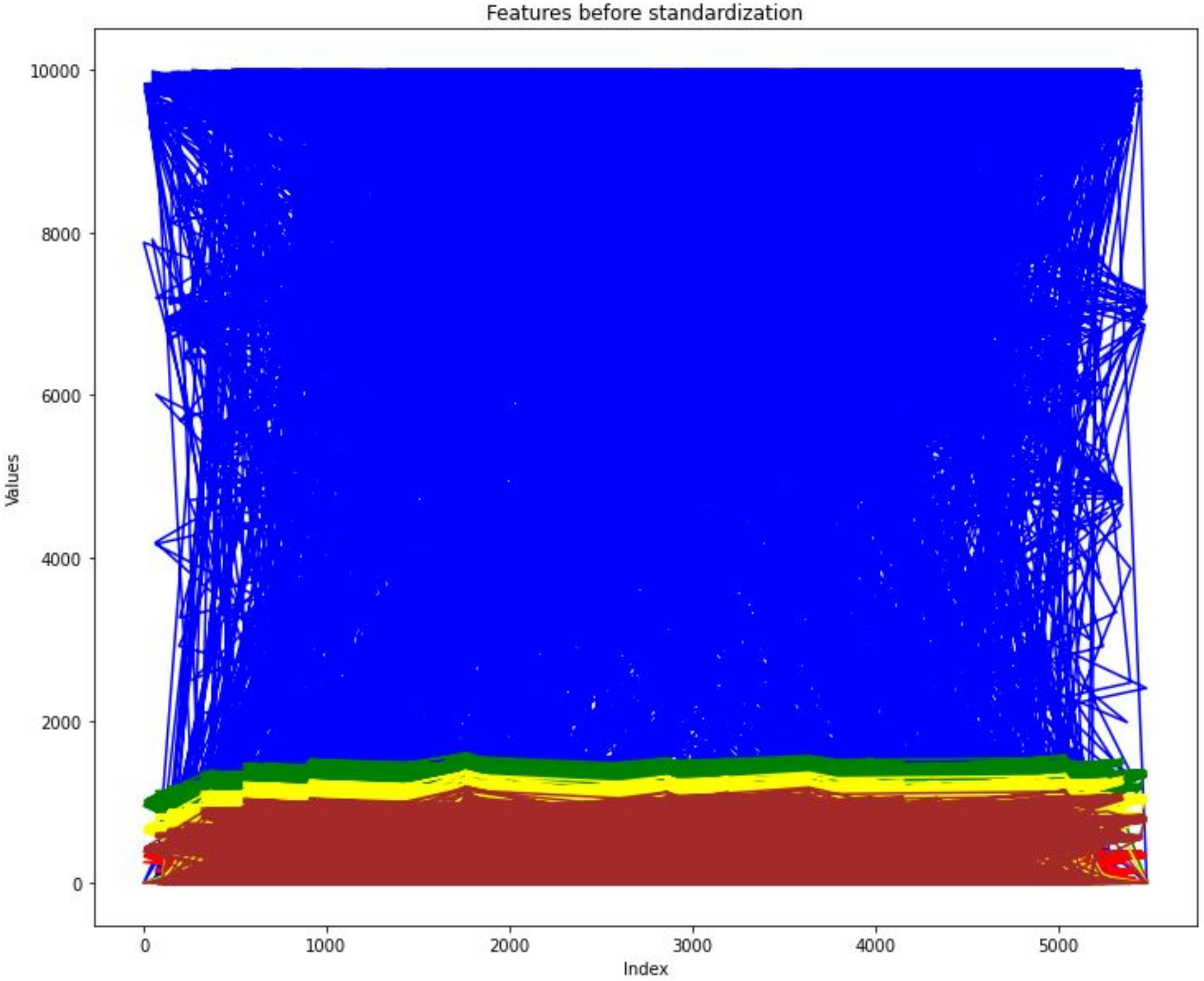
	MONTH	DAY	T_AMB	P_AMB	CMP_SPEED	CDP	GGDP	HPT_IT	CDT	LPT_IT	EXH_T	RH	WAR	POWER
0	1	1	1.450440	0.843522	0.000000	0.843522	0.843522	0.000000	1.450440	0.000000	0.000000	81.237441	0.000041	0.000000
1	1	2	2.761142	0.843856	7870.729713	7.907587	2.448490	949.263690	258.933367	625.677722	387.749872	74.311313	0.000041	13332.692409
2	1	3	9.270325	0.843413	9898.625866	9.407523	2.816769	984.601577	338.014765	655.857137	413.039467	47.897182	0.000041	13026.684965
3	1	4	14.293265	0.844249	9850.791469	9.121784	2.775070	1014.536922	347.129100	681.701087	434.895488	34.400729	0.000041	12773.507042
4	1	5	12.875213	0.843663	9828.508458	9.138088	2.776577	1008.503746	344.360211	677.018748	431.268990	37.537882	0.000041	12768.092781

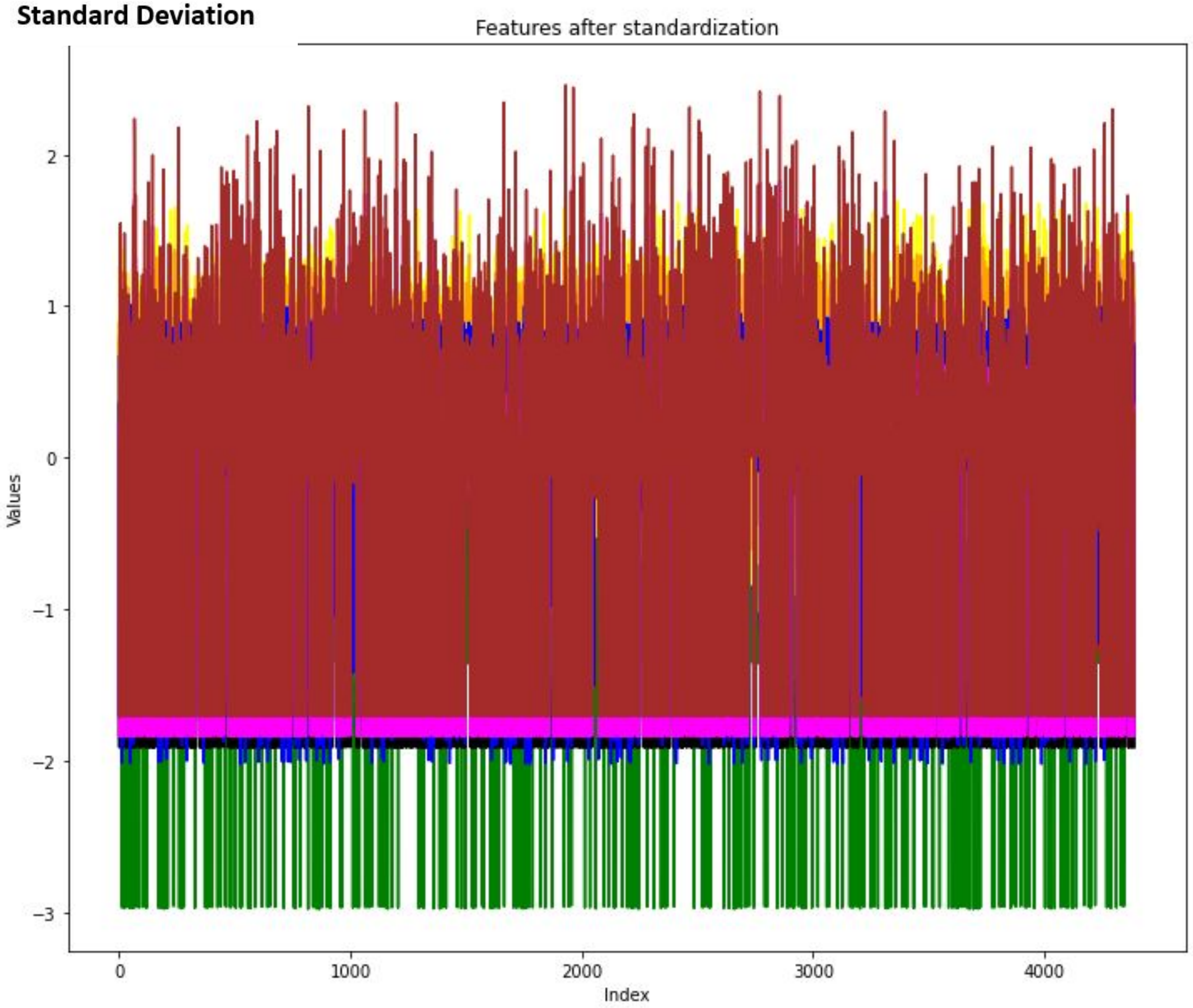


	P_AMB	CMP_SPEED	CDP	GGDP	HPT_IT	CDT	LPT_IT	EXH_T	POWER
0	0.843522	0.000000	0.843522	0.843522	0.000000	1.450440	0.000000	0.000000	0.000000
1	0.843856	7870.729713	7.907587	2.448490	949.263690	258.933367	625.677722	387.749872	13332.692409
2	0.843413	9898.625866	9.407523	2.816769	984.601577	338.014765	655.857137	413.039467	13026.684965
3	0.844249	9850.791469	9.121784	2.775070	1014.536922	347.129100	681.701087	434.895488	12773.507042
4	0.843663	9828.508458	9.138088	2.776577	1008.503746	344.360211	677.018748	431.268990	12768.092781

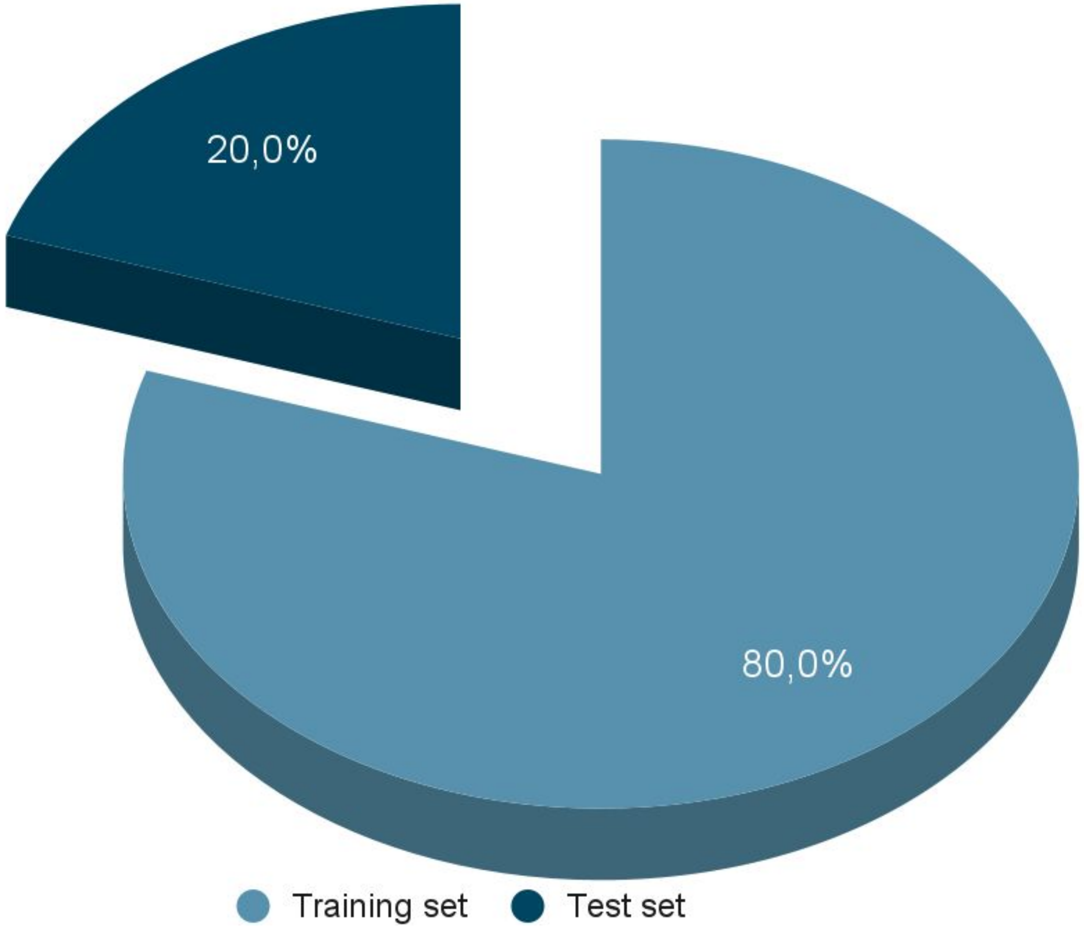
Data Transformation

$$X_{new} = \frac{X_i - X_{mean}}{\textbf{Standard Deviation}}$$

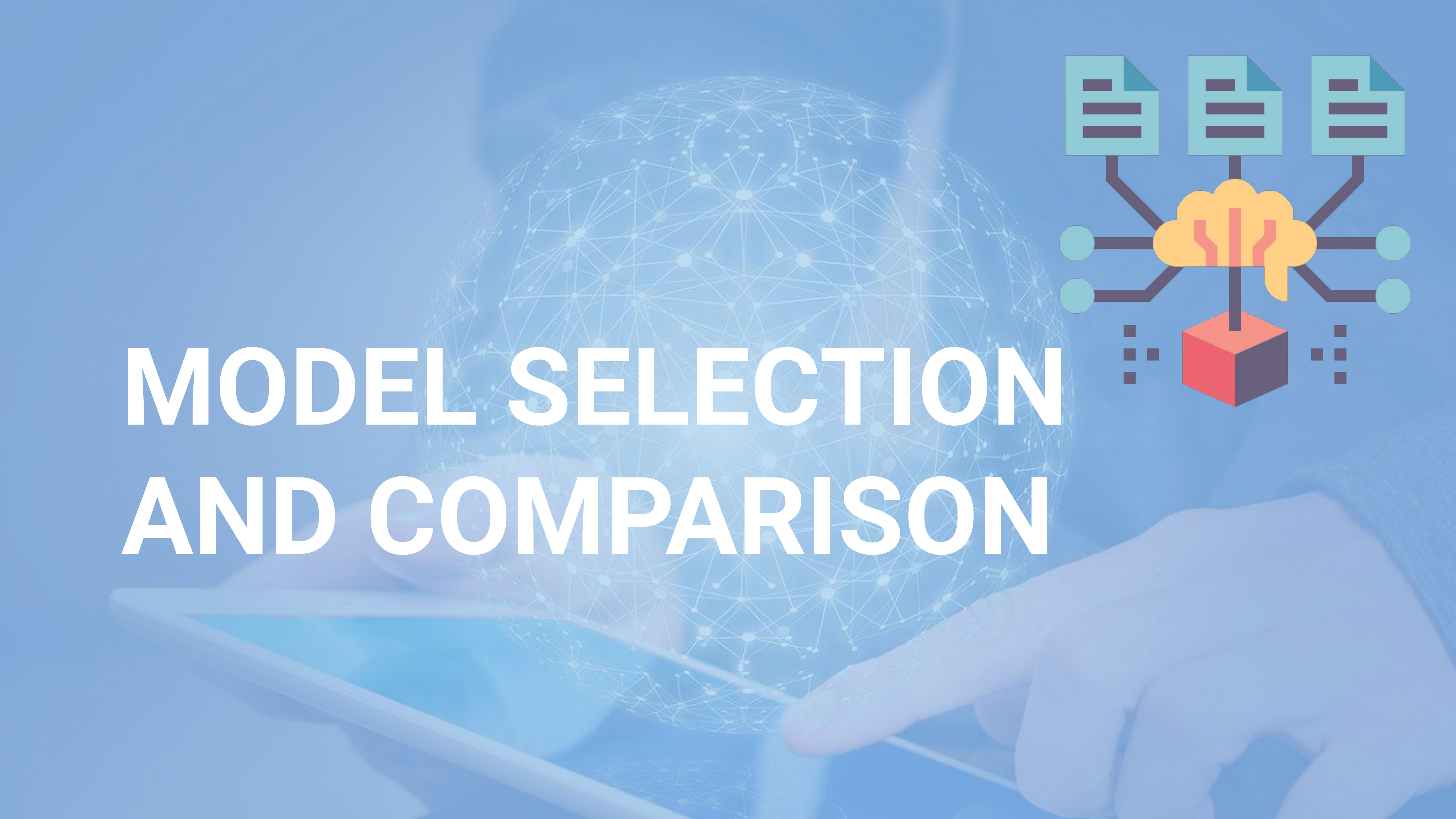




Data partitioning



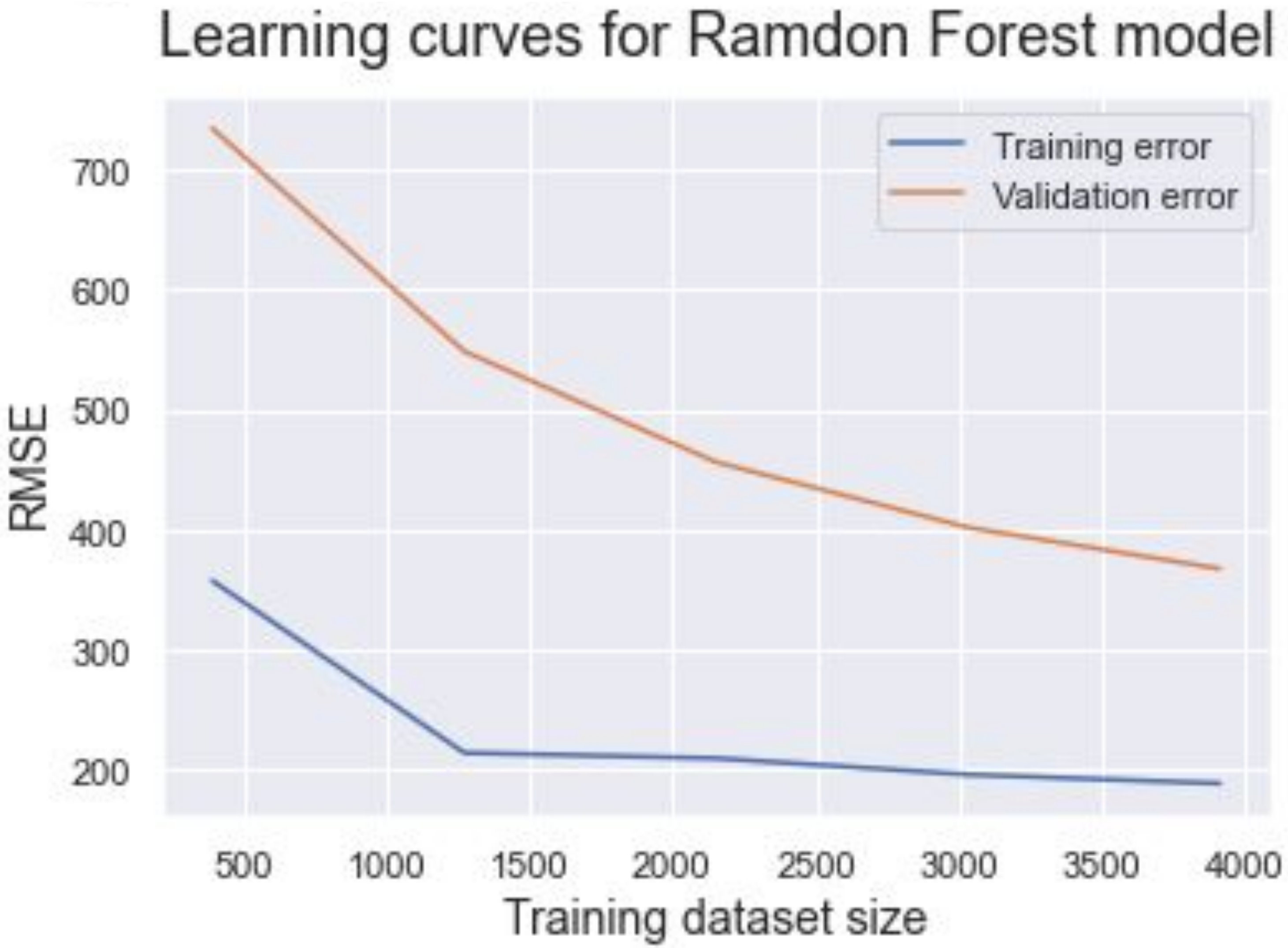
MODEL SELECTION AND COMPARISON



Where to start?

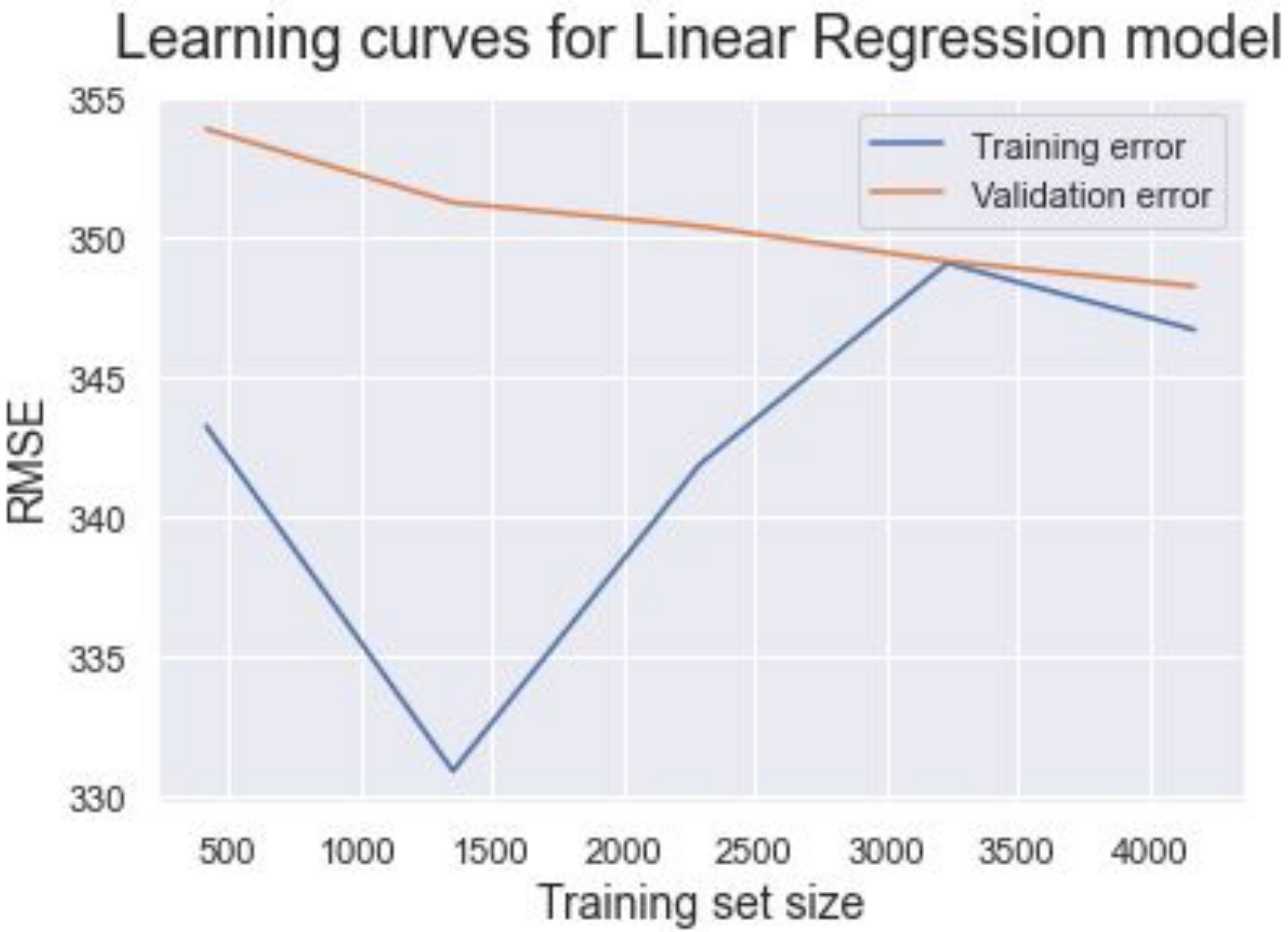
- Is it a classifier or regression?
- Does it have a linear behavior?
- How big is the data set?

Random Forest



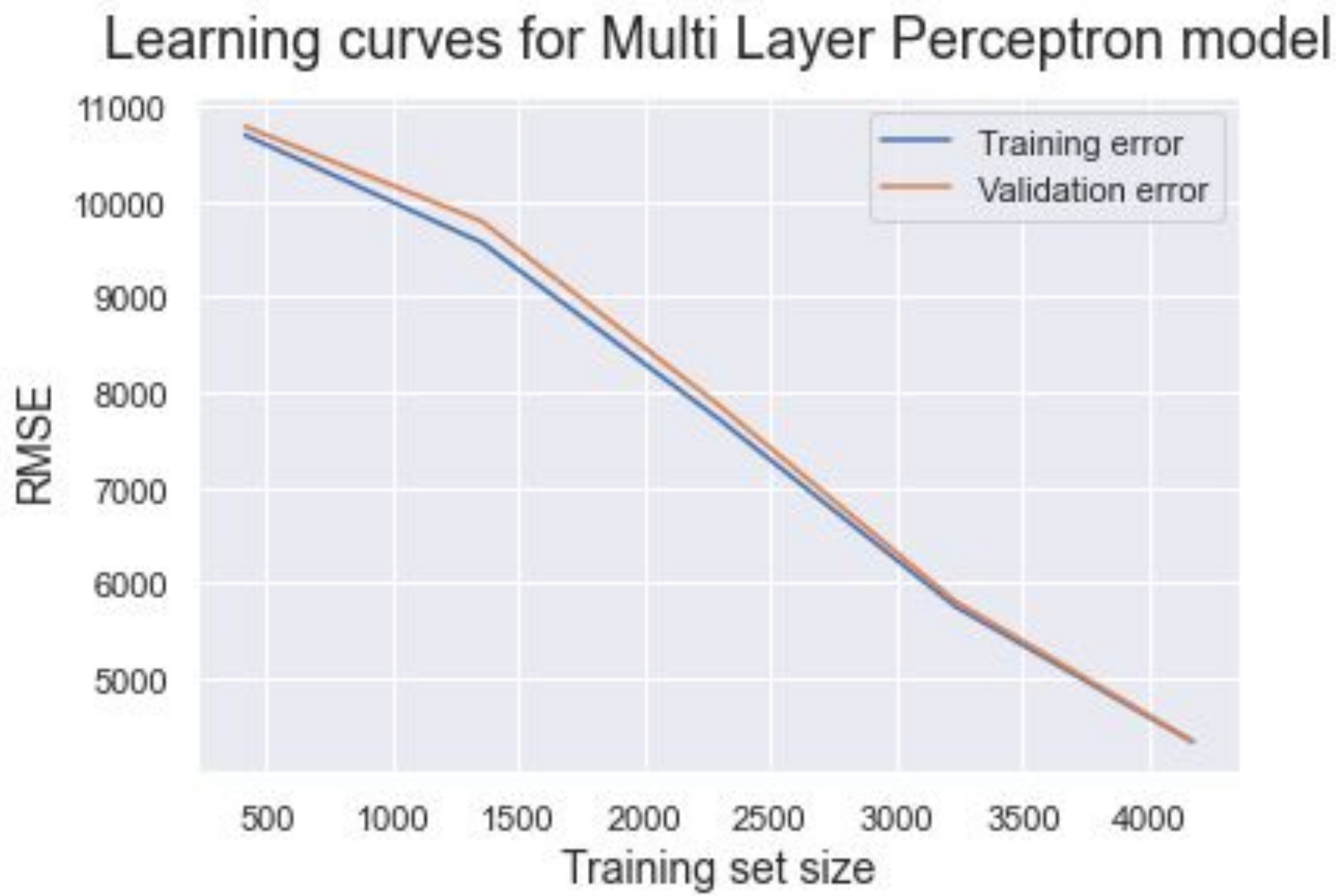
R2 = 0.9949

Linear Regression

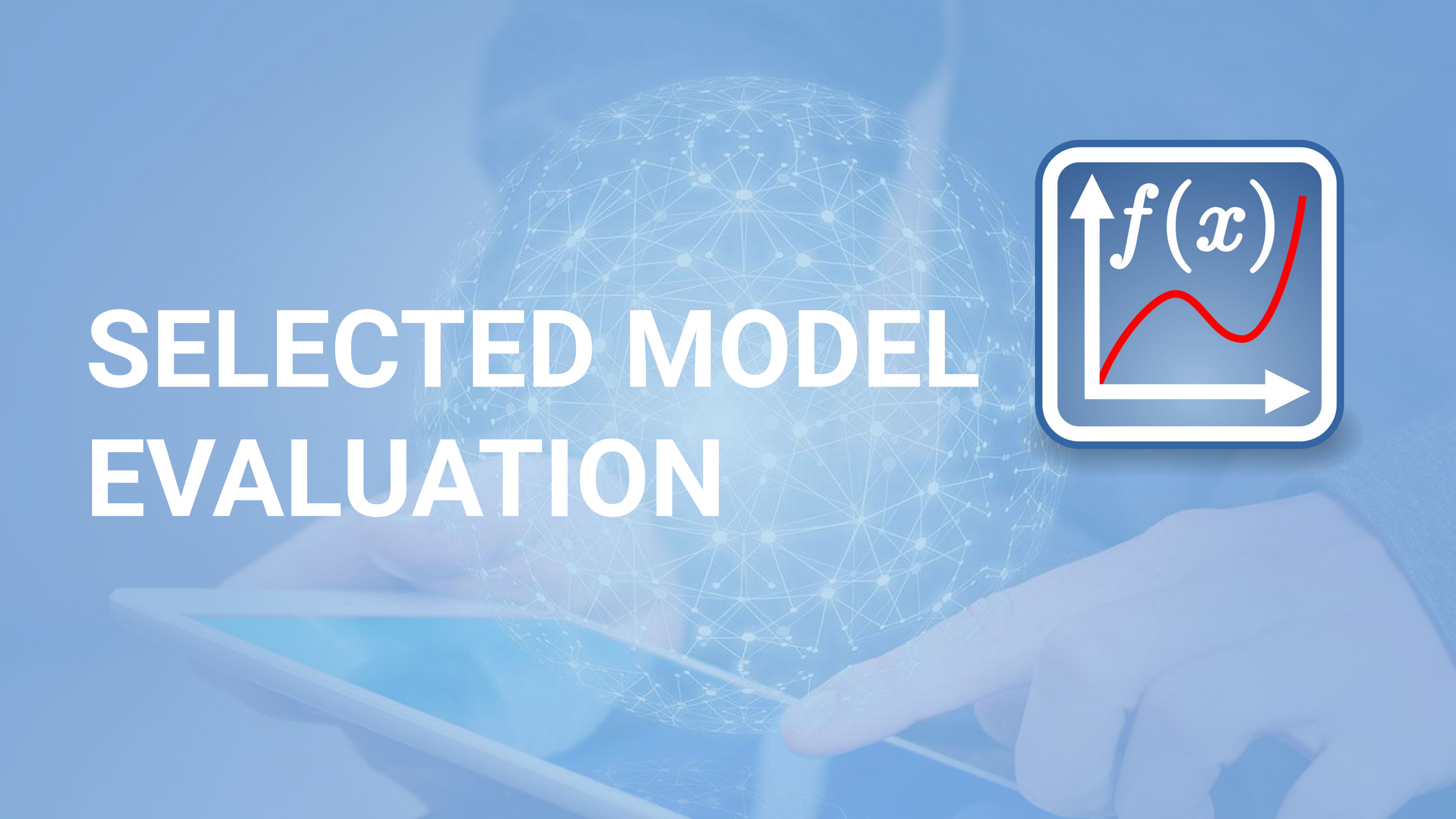


R2 = 0.9963

Neural Network



R2 = 0.5369



SELECTED MODEL EVALUATION

Training selected model

Linear Regression takes it all

Model Parameters:

- **fit_intercept=True**
- **normalize='deprecated'**
- **copy_X=True**
- **n_jobs=None**
- **positive=False**

Model Attributes:

coef_ = Estimated coefficients for the linear regression problem

intercept_ = Independent term in the linear model.

Validation Results

R2 = 0.9964

RMSE = 354.343

Testing with Kaggle

RMSE= 320.24

Final model

$$POWER = 9275.5 + 4.32e^{1}(p\_amb) - 1.77e^{3}(cmp\_speed) - 2.19e(cdp) + 2.73e^{3}(ggdp) + 1.32e^{5}(hpt\_it) - 7.17e^{2}(cdt) - 2.18e^{5}(lpt\_it) + 9.14e^{4}(exh\_t)$$

Conclusions

ADJUST OF DATA PERFORMANCE

DIFFERENCE WITH OTHER MODELS